基于卷积神经网络的图像隐写分析方法 *

高培贤 a, b，魏立线 a, b，刘　佳 a, b，刘明明 a, 2

(武警工程大学 a. 网络与信息安全武警部队重点实验室; b. 电子技术系，西安 710086)

摘　要：为了提高卷积神经网络(CNN)在图像隐写分析领域的分类效果，构建了一个新的卷积神经网络模型（steganalysis-convolutional neural networks，S-CNN）进行隐写分析。该模型采用两层卷积层和两层全连接层，减少了卷积层的层数；通过在激活函数前增加批量正规化层对模型进行优化，避免了模型在训练过程中陷入过拟合；取消池化层，减少嵌入信息的损失，从而提高模型的分类效果。实验结果表明，相比传统的图像隐写分析方法，该模型减少了隐写分析步骤，并且具有较高的隐写分析准确率。

关键词：图像隐写分析；卷积神经网络；批量正规化；激活函数

中图分类号：P309.2　　doi: 10.3969/j.issn.1001-3695.2017.07.0692

Image steganalysis based on convolution neural network

Gao Peixian[a, b], Wei Lixian[a, b], Liu Jia[a, b], Liu Mingming[a, b]

(*a. Key Laboratory for Network & Information Security of Chinese Armed Police Force*, *b. Dept. of Electronic Technology*, *Engineering University of Chinese Armed Police Force*, *Xi'an* 710086, *China*)

Abstract: In order to improve the recognition effect of convolutional neural networks(CNN) in image steganalysis, this paper constructed a new steganalysis-convolutional neural networks model （S-CNN）for steganalysis. The model reduced the number of layers of the convolution layer by using two layers of convolution layer and two layers of the whole connection layer. By adding the batch normalization layer to optimize the model before the activation function, to avoid the model in the training process into the over-fitting. The cancellation of the pool layer reduced the loss of embedded information, thereby improving the classification effect of the mode. The experimental results show that, compared with the traditional steganalysis methods, the proposed model reduces the steganalysis step and has higher steganalysis accuracy.

Key Words: image steganalysis; CNN; batch normalization; activation function

0　引言

目前，图像隐写分析成为信息安全领域研究热点[1,2]之一。传统的图像隐写分析方法分为两步：一是特征提取，例如小波直方图特征、马尔科夫特征、离散余弦变换系数的共生矩阵特征[3,4]。二是特征分类，例如 Fisher 算法、支持向量机[5,6]。由于传统的方法可靠性低或者是训练过程非常耗时，会对隐写分析效率造成不利影响。

近几年，随着深度学习在图像分类领域优异的表现，越来越多的研究者将深度学习应用在自己的领域。2015 年，Qian 等人使用深度学习来代替传统的两步机器学习进行隐写分析[7]，提出了五层的 CNN 模型，该模型使用高通滤波器（HPF）进行卷积预处理，激活函数采用 Gaussian 非线性函数，池化层使用的是平均池化。通过在 BOSSbase 数据集上测试，获得的准确率仅仅比空域富模型（SRM）[8]+集成分类器（EC）[9]低 3%到 4%。这个结果对于图像隐写分析来说是一个巨大的进步。2016 年 5 月，Guanshuo Xu 等人提出了一个五层的 CNN 模型，该模型在第一个卷积层后加入了一个绝对值层（ABS）来增强网络后边卷积层的学习能力。为了避免过拟合，他们在前几层约束了数据的范围并且在更深层次采用大小 1×1 的卷积核。在透视场景下，对 S-UNIWARD 隐写算法的检测准确率达到 80.24%[10]。这些结果都表明了 CNN 在图像隐写分析领域具有巨大的潜力。

为了提高图像隐写分析的正确率和可靠性，本文构建了一个两层的 CNN 模型进行图像隐写分析。实验结果表明，相对于 Guanshuo Xu 等人的 CNN 模型，本文模型进行图像隐写分析的检测准确率提高了 8.68%。

基金项目：国家自然科学基金资助项目(61403417)；国家重点研发计划资助项目(2017YFB0802002)；陕西省自然科学基础研究计划资助项目(2016JQ6037)

作者简介：高培贤（1994-），男山西临汾人，硕士研究生，主要研究方向为信息隐藏（1198074954@qq.com）；魏立线（1966-），男，教授，硕士，主要研究方向为信息安全；刘佳（1982-），男，讲师，博士，主要研究方向为模式识别、图像隐写分析；刘明明（1991-），男，硕士研究生，主要研究方向为信息隐藏.

1　五层 CNN 框架

Guanshuo Xu 等人构建的模型由五组卷积模块组成，如图 1 所示，其中流程图框内的算式为：（卷积核个数）×（卷积核的高×卷积核的宽×输入特征图数），框下方的算式为：（特征图数）×（图像的高×图像的宽），卷积层和池化层以及 HPF 层都使用了填充。该模型每组卷积模块由卷积层提取特征开始，到平均池化层结束（第五组为全局池化）。第 1 组和第 2 组的激活函数采用的是 TanH 激活函数，其余的为 ReLU 激活函数。为了约束数据的范围，在第一组卷积模块采用了绝对值层（ABS）。为了避免 CNN 训练陷入局部最优化，在每个激活函数层加入了批量正规化层（Batch Normalization Layer，简称 BN 层）。输出层包括一层全连接层和一层损失函数层。该模型在 BOSSBass 数据库上检测 S-UNIWARD 嵌入算法的准确率达到 80.24%。

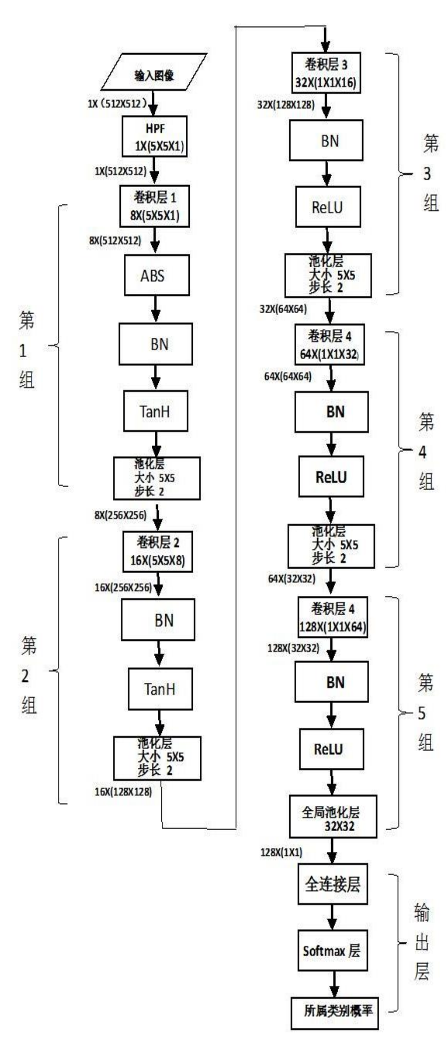


图 1　Guanshuo Xu 等人构建的 CNN 模型

2　S-CNN 框架

基于 Guanshuo Xu 等人的研究，本文构建了一个新的 CNN 模型进行图像隐写分析，如图 2 所示，　整个 S-CNN 框架分为输入模块，卷积模块和输出模块。首先，原始图像经过高通滤波层（HPF 层）进入卷积模块，在卷积模块中进行卷积运算提取特征，最后经过输出模块输出该图像所属类别的概率，完成分类。

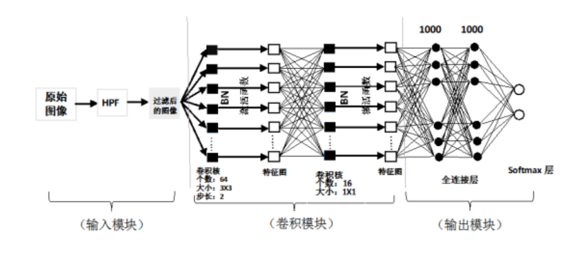


图 2　本文构建的 S-CNN 模型

对于图像隐写分析，进入网络的图像首先经过 HPF 预处理，HPF 层可以加快 CNN 模型的收敛速度。HPF 层是一种特殊的卷积层，HPF 层的卷积核大小为 5x5，权重初始值为 F。通过设置 HPF 层的学习率参数为 0 使得权重 F 固定不随着训练更新。

$$F=\frac{1}{12}\begin{pmatrix}-1&2&-2&2&-1\\2&-6&8&-6&2\\-2&8&-12&8&-2\\2&-6&8&-6&2\\-1&2&-2&2&-1\end{pmatrix}$$

卷积模块由卷积层，激活函数层组成。卷积层通过卷积计算从而提取出原始图像的不同特征。原始图像都存在局部稳定性，卷积层把各区域的特征综合起来就可以得到整幅图像的特征。卷积层由多个卷积核组成，每个卷积核是一个权值矩阵，卷积核越多提取的特征就越多。如图 2，经过 HPF 过滤后的图像与卷积核进行矩阵运算，提取出输入图像的不同特征，每个卷积核产生一个特征图，该特征图作为激活函数层的输入，激活函数的使用打破了在卷积过程中进行线性滤波的线性特性。图像隐写是对载体图像中嵌入噪声，图像隐写分析是识别图像中的噪声，池化层会弱化这种噪声，对分类效果产生消极的影响。因此，本文取消了池化层。

经过两组卷积模块后，特征图进入输出模块。输出模块包括两层全连接层和损失函数层。全连接层中的每个神经元都与其前一层的所有神经元连接。全连接层可以整合卷积层或者取样层中具有类别区分型的局部信息[11]。为了提升 CNN 网络性能，全连接层的每个神经元的激活函数采用了 ReLU 函数。损失函数层的损失函数采用的是 softmax 函数，对于具体的分类问题，选择合适的损失函数至关重要。最后由 softmax 层完成分类。

2.1　激活函数

激活函数的选择是构建 CNN 过程中的重要环节。本文构建的 CNN 模型的激活函数为 ReLU 函数，并通过实验对比了 TanH 函数和 ReLU 函数对分类效果的影响。图 3 中实线为 ReLU 曲线，虚线为 TanH 曲线，从图中可以看出使用 ReLU 函数时，函数的输出不会随着输入的增加而趋于饱和。相比饱和非线性

函数，非饱和非线性函数能够解决梯度爆炸/梯度消失问题，同时也能加快收敛速度[12]。为了计算反向传播回来的误差，激活函数必须可导。

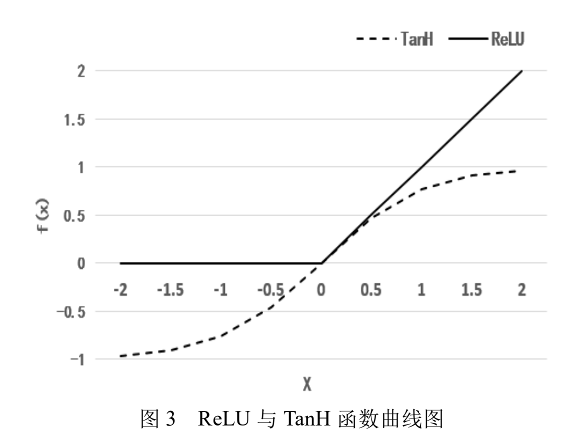


图 3　ReLU 与 TanH 函数曲线图

2.2　批量正规化

对于神经网络来说，网络的训练是一个复杂的过程，除了输入层的数据外，随着网络参数的更新，后面每一层网络的输入数据分布始终在发生着变化，这种变化会在网络中累计放大下去，导致网络需要重新学习新的数据分布，进而影响网络的训练速度，这种数据分布的改变，一般称之为“Interal Covariate Shift” [13]。另一方面，神经网络学习过程的本质就是学习数据分布，一旦训练数据与测试数据的分布不同，那么网络的泛化能力也大大降低。为了提高训练效率和识别效果，本文在激活函数前添加批量正规化层（BN 层）。BN 算法分为两步：第一步是在每一层的输入前进行归一化处理，将上一层的输出数据归一化至：均值为 0、方差为 1。其公式表达如下：

$$\hat{x}^{(k)} = \frac{x^{(k)} - E\left[x^{(k)}\right]}{\sqrt{Var\left[x^{(k)}\right]}}$$

其中：K 表示网络的层数，$x^{(k)}$ 表示第 K 层网络中 BN 层的输入图像。注意：$E\left[x^{(k)}\right]$ 为训练中每一批次（mini-batch）的图像的平均值，$Var\left[x^{(k)}\right]$ 为训练中每一批次（mini-batch）的方差，而不是全体数据集。第一步的归一化会破坏上一层网络学到的特征，为了恢复上一层网络所学习的特征，BN 算法第二步是进行变换重构。其公式表达如下：

$$y^{k} = \gamma^{k} \hat{x}^{(k)} + \beta^{k}$$

其中：y^{k} 为 BN 层的输出图像，γ、β 为可学习的参数，当 $\gamma^{k} = \sqrt{Var\left[x^{(k)}\right]}$，$\beta^{k} = E\left[x^{(k)}\right]$ 时，便可恢复上一层网络所学习到的特征。通过变换重构，特征图的空间关系得以保存。

2.3　模型分析

相比 Guanshuo Xu 等人的模型，本文构建的 S-CNN 减少了 CNN 的层数，增加了卷积层的卷积核的个数，从而使得网络提取到更多的特征。池化层可以减少网络的计算量，取消池化层会增加内存消耗，但是池化层会导致信息的损失，不利于隐写分析，综合考虑，本文取消了池化层。本文的激活函数全部为 ReLU 激活函数，在实验中也证明了激活函数全为 ReLU 函数的识别效果好于第一层为 TanH 函数和第二层为 ReLU 函数。另外，本文构建的模型全连接层为两层，这也是识别效果有所提升的原因之一。BN 层的使用避免了模型在训练过程中陷入局部最优化，并且提高了识别准确率，因此，本文的 S-CNN 模型在卷积层之后使用了 BN 层。

3　实验

3.1　数据集以及实验平台

实验中的数据集采用的是 BOSSbase V1.01。该数据集中含有 10 000 张分辨率为 512X512 的 pgm 格式的灰度图像，是目前进行图像隐写分析研究的最为常用的数据集。由于 GPU 显存的限制，在实验中利用 Photoshop 的批命令将数据集中的灰度图像裁剪至大小为 128X128 的图像，并随机取其中的 40 000 张作为载体图像。用 S-UNIWARD 隐写算法对载体图像进行嵌入，嵌入率为 0.4 bpp。本文将 30 000 张载体图像和 30 000 张隐写图像作为训练样本。10 000 张载体图像和 10 000 张隐写图像作为测试样本。实验是在 windows 下用开源的深度学习框架 caffe 实现的。

3.2　实验过程

根据 caffe 平台的要求，实验中先将数据集转换为 leveldb 格式，并计算样本的均值。在网络结构文件中搭建 CNN 进行训练，并且保存训练日志用于性能分析。为了提高 CNN 网络的训练效率，在图像预处理阶段对原始图像进行归一化处理。每张原始图像都减去图像集的均值可以极大的减小计算机计算资源消耗，加快训练速度。移除 drop out、L2 正则项参数，提高网络模型的泛化能力[13]。BN 层的使用可以使得模型选择较大的学习效率，从而大大加快模型收敛速度，在相同的学习效率情况下，BN 层的使用也可以减少训练所需要的时间，加快模型的收敛速度。为了提高识别准确率，打乱了数据集中的原始图片顺序，防止原始图像集中的某一图像被多次调用。

3.3　实验参数

Caffe 提供六种优化算法，在实验中本文选择的是随机梯度下降算法（SGD）。基础学习率（base_lr）为 0.01，上一次更新的权重（momentum）为 0.9，权值衰减（weight_decay）为 0.004。学习率策略（lr_policy）选择的是”inv”，学习率会随着迭代次数的增加而减小，避免了手动调整学习率参数。由于显存的限制，训练时每一批次（batch size）为 64。最大迭代次数为 5000，为了更好的评估 CNN 模型的性能，每迭代 300 次进行一次测试，共测试 16 次。对于 HPF 层，学习率（lr_mult）设为 0，使其参数固定不更新。

3.4　实验结果

相比传统的 CNN 隐写分析方法，S-CNN 模型减少了网络

的层数，简化了网络结构，从而加快了模型的训练效率，同时又提高了模型的识别准确率。在相同的实验设备下，各模型的训练所需时间以及识别准确率如表 1 所示。

表 1　各模型训练时间以及识别准确率

模型	训练时间（h）	准确率
Tan 等人的 3 层 CNN 模型[14]	2.7	69%
Qian 等人的 5 层 CNN 模型[7]	3.5	76.19%
Xu 等人的 5 层 CNN 模型[10]	3.2	80.24
S-CNN 模型	2	88.92%

相比传统的两步机器学习隐写分析法（RM+EC），S-CNN 模型对 S-UNIWARD 隐写算法的检测效果也有了较大的提升。实验结果如表 2 所示。

表 2　各模型识别准确率对比结果

模型	准确率
RM+EC	79.53%
S-CNN 模型	88.92%

根据 GPU 的性能，综合训练时间和识别准确率的考虑，本文测试了五类卷积核大小不同的 CNN 网络，结果取 16 次测试结果的平均值。为了验证 BN 层对识别准确率的影响，本文还进行了有无 BN 层的对比实验，结果证明 BN 层可以有效提高 CNN 模型的识别效果，如表 3 所示。

表 3　不同卷积核大小的 CNN 的识别准确率

卷积层 1	卷积层 2	准确率（无 BN 层）	准确率（有 BN 层）
7X7	5X5	0.8443	0.8584
6x6	4X4	0.8534	0.864
5X5	3X3	0.8539	0.8652
4X4	2X2	0.8608	0.8724
3X3	1X1	0.8773	0.8892

本文还对激活函数的选择做了对比实验，实验结果表明，激活函数全部选择 ReLU 函数的效果优于第一层选择 TanH 函数和第二层 ReLU 函数。结果如图 4 所示。

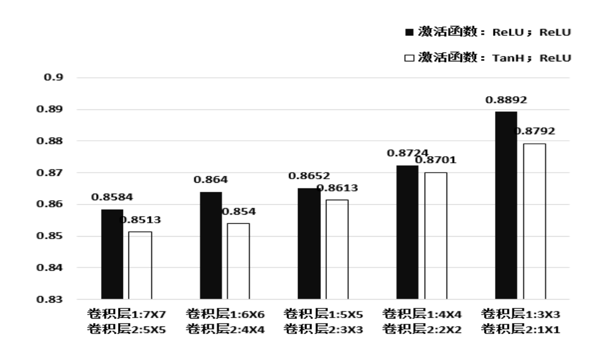


图 4　不同激活函数的模型的识别准确率

对于图像隐写分析领域，HPF 层可以极大的提高模型的收敛速度，表 4 为本文最优模型训练迭代 5 000 次时有无 HPF 层的损失值和准确率。

表 4　有无 HPF 层的损失值（loss）和准确率

模型	损失值	准确率
有 HPF 层	0.0148	0.8892
无 HPF 层	6.4572	0.5012

损失值（loss）反映的估计值和真实值之间的误差，代表着模型的拟合程度。图 5 为本文构建的模型训练过程中的 loss 值变化情况。随着迭代次数的增加，loss 值呈递减趋势并趋于稳定。

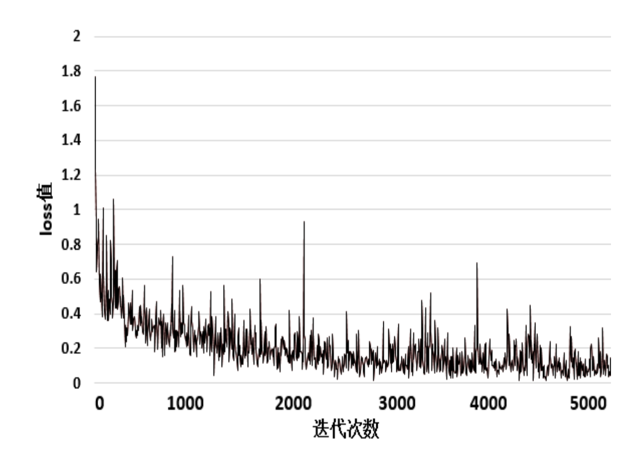


图 5　训练过程中的损失值变化情况

目前生成对抗网络（GAN）在图像隐写分析领域的研究已成为当前的热点之一。GAN 通过生成网络和判别网络互相博弈，从而达到一种平衡。Denis 等人[15]提出的 DCGAN 模型的中的分类器识别准确率如表 5 所示。

表 5　DCGAN 与 SCNN 的识别准确率对比

模型	准确率
GAN（相同种子）	0.982
GAN（随机种子）	0.649
S-CNN	0.8892

从表中可以看出，当生成网络的种子相同时，识别效果优于 S-CNN 模型，但生成的图像不利于图像隐写。当生成网络的种子随机时，识别效果降低，远低于 S-CNN 模型。但对于图像隐写来说可以产生更好的载体图像。

4　结束语

实验结果表明，深度学习是一种非常有前景的图像隐写分析工具，相比传统的图像隐写分析工具，避免了人工提取特征，大大提高了图像隐写分析的效率，并且有效提高了隐写分析的准确率。目前，CNN 只能对简单的隐写算法进行测试，当算法复杂或者嵌入率过低时，分类效果会降低。今后的工作主要围绕着增强 CNN 隐写分析的通用性进行。

参考文献：

[1] 任洪娥，常春武，张健. 改进的双线性插值算法在信息隐藏中的应用[J]. 计算机应用研究, 2010, 27 (11): 4290-4292.

[2] 陶然，张涛，平西建. 基于纹理复杂度和差分的抗盲检测图像隐写算法[J]. 计算机应用, 2011, 31 (10): 2678-2681.

[3] Kodovsky J, Fridrich J, Holub V. Ensemble classifiers For steganalysis of digital media [J]. IEEE Trans on Information Forensics & Security, 2012, 7 (2): 432-444.

[4] Luo W, Huang F, Huang J. Edge adaptive image steganography based on LSB matching revisited [J]. IEEE Trans on Information Forensics & Security, 2010, 5 (2): 201-214.

[5] Pevný T, Bas P, Fridrich J. Steganalysis by subtractive pixel adjacency matrix [J]. IEEE Trans on Information Forensics & Security, 2010, 5 (2): 215-224.

[6] Xiong Gang, Ping Xijian, Zhang Tao, et al. Image textural features for steganalysis of spatial domain steganography [J]. Journal of Electronic Imaging, 2012, 21 (3): 033015.

[7] Qian Y, Dong J, Wang W. Deep learning for steganalysis via convolutional neural networks [C]// Proc of International Society for Optical Engineering. 2015: 9409: 94090J-94090J-10.

[8] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images [J]. IEEE Trans on Information Forensics & Security, 2012, 7 (3): 868-882.

[9] Kodovsky J, Fridrich J, Holub V. Ensemble classifiers for steganalysis of digital media [J]. IEEE Trans on Information Forensics & Security, 2012, 7 (2): 432-444.

[10] Xu G, Wu H Z, Shi Y Q. Structural design of convolutional neural networks for steganalysis [J]. IEEE Signal Processing Letters, 2016, 23 (5): 708-712.

[11] Sainath T N, Mohamed A R, Kingsbury B, et al. Deep convolutional neural networks for LVCSR [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 8614-8618.

[12] Xu Bing, Wang Naiyan, Chen Tianqi, et al. Empirical evaluation of rectified activations in convolution network [EB//OL]. arXiv: 1505. 00853v2, 2015.

[13] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [Z]. 2015.

[14] Tan S, Li B. Stacked convolutional auto-encoders for steganalysis of digital images [C]// Proc of IEEE Signal and Information Processing Association Summit and Conference. 2014: 1-4.

[15] Volkhonskiy D, Nazarov I, Borisenko B, et al. Steganographic generative adversarial networks [Z]. 2017.